Subjective Duration Assessment: An Implicit Probe for Software Usability

Mary Czerwinski, Eric Horvitz, Edward Cutrell

Microsoft Research
One Microsoft Way
Redmond WA 98052 USA
{marycz ;horvitz ;cutrell}@microsoft.com

SUMMARY

This paper explores a new approach to gauging users' difficulties with tasks, interfaces, and situations we refer to as subjective duration assessment. The approach, adapted from results described in the interruption literature in psychology, centers on the use of time estimation to characterize performance. We introduce a metric, named relative subjective duration (RSD) that provides a means for probing the difficulty that users have with performing tasks—without directly asking users about the difficulty We do this in order to avoid the inherent bias toward the positive end of the scale typically seen in user satisfaction ratings after usability studies. It has been shown previously that when engaging tasks are interrupted, participants tend to overestimate how long those tasks take when compared to actual task times. Conversely, tasks that are completed tend to be underestimated in terms of the overall task times. We have explored the value of time estimation as a metric for evaluating task performance in HCI. Our hypothesis was that the duration of activity on tasks that are halted before completion would be overestimated, because participants were not able to complete them on their own, while the duration of activity on tasks completed successfully would be underestimated. A user study of a well-known Internet browser explored the efficacy of the metric. Our results show that within deployment constraints, RSD can be a valuable tool for HCI research.

KEYWORDS: Usability, time perception, time estimation, metrics, empirical studies

INTRODUCTION

A classic problem in the usability engineering discipline is the problem of how to interpret study results when performance on a user interface is poor but user satisfaction with the design is relatively high. It is well known that participants in usability studies often provide user satisfaction measures that are more positive than would normally be expected (e.g., [3]). It is our premise in this paper that asking users to perform time estimations for tasks during usability studies might be an implicit means for ascertaining a less biased measure of deliberative effort with tasks. This could be invaluable during usability studies in terms of making real sense out of performance and more qualitative metrics, and would be a definite contribution to the field of usability engineering. We thought it would be interesting to see how time estimation as a dependent measure varied with task completion and task difficulty, as well as satisfaction. We present an initial user study that examines this dependent measure during a standard web usability study. Our results suggest that time estimation is a valuable implicit metric for the field of HCI. This initial examination suggests that time estimation can provide a unique and powerful combination of subjective and performance-based data for a wide range of usability studies.

One of the earliest uses of time estimation in psychological experiments can be traced back to a phenomenon subsequently termed the *Ziegarnik effect*. Ziegarnik [6] ran a large set of studies wherein participants were given different tasks to perform. Prior to completing some of these tasks, participants were instructed to terminate working on that task and to switch to something else. Ziegarnik found that participants' abilities to freely recall tasks performed during a typical session showed a reliable advantage for uncompleted over completed tasks. This result was replicated many times over the years. Bergen (1968) reviewed approximately 40 years of interruption research and theory in psychology, including an interpretation of the original memory effects labeled "Ziegarnik" effects as well as extensions and replications of the original studies. Van Bergen noted that subjects in Ziegarnik-like studies typically remembered items from uncompleted tasks better than completed tasks, *as long as the tasks were engaging and subjects were motivated by the instructions.*

TIME ESTIMATION

Weybrew [7] extended the Ziegarnik effect to the realm of time estimation. Weybrew studied the perceived length of time of 2 different kinds of tasks with and without interruption. The two task categories were math problems (the addition of random 3 digit numbers) and letter cancellation (less difficult, canceling i's and s's in text). The tasks were practiced, and then begun again, following a break. After resumption, half of the tasks were interrupted, and participants were not allowed to

complete the interrupted task. Participants then estimated how long each of the 2 phases took. They were challenged with arithmetic problems first and then were transferred to the letter cancellation task. Task length estimates were converted into percentages—(time estimated-time total)/time total—and analyzed. Results revealed borderline reliable findings for interrupted tasks being overestimated, and for non-interrupted tasks being underestimated. This finding was very similar to the memory finding in the original Ziegarnik studies. Weybrew found that the letter cancellation task (which was quite boring and repetitious) was more underestimated when not interrupted, but not significantly so. He also found that a borderline reliable effect (p-value was .06 with a small sample size) for the more difficult, arithmetic tasks strongly overestimated when interrupted. So, Weybrew replicated findings summarized by Van Bergen that the more engaging or difficult the task, the stronger the Ziegarnik effect, albeit focusing on time estimation as opposed to a free recall dependent measure.

Upon review of many current psychological papers in the field of psychology with regard to time estimation, it appears that time estimates are accumulated with influence from multiple channels of information, with some channels influencing perception more heavily than others, due to their relatively lower need for attentional resources [1, 2, 4]. In addition, interruptions that require more resources due to their difficulty or attentional demands may disrupt visual input into some kind of time estimation "accumulator" more than auditory input. Depending on the perceptual channel most engaged in interrupted task performance, then, one may or may not see under or overestimation of task time relative to actual task time.

USER STUDY—WEB BROWSER USABILITY

Based on the Weybrew adaptation of the Ziegarnik effect, we believed that if participants were unsuccessful at a given task during the course of a usability study, they might overestimate how long that task took. If participants were engaged and successful with the task, they might underestimate task length. Our interest in this metric stemmed from its well-established acceptance in the field of psychology and its implicit nature (participants probably do know what we were measuring when we ask for time estimates). To that end, we recruited six novice-to-intermediate experienced Internet users, aged 31-55, to participate in this study. Participants were screened using an internal, well-validated screening tool for Internet expertise.

METHODS AND PROCEDURE

A standard usability study was run as an iterative test of a well known, Internet browser, including seventeen typical Internet browser tasks, such as logging in, account maintenance, playing videos and songs, searching, sending instant messages, email and calendar activities on the web. Task success rates with and without experimenter intervention, completion times, participants' estimates of how long each task took, and overall user satisfaction ratings were collected as dependent measures. If a participant was in an error state during a given task's execution for more than two minutes, or if the participant explicitly either verbally "gave up" or asked for the experimenter's assistance, this task was considered a failure without experimenter intervention. However, the experimenter would, after varying levels of intervention (ranging from a hint to an explicit solution to the task problem) allow the user to continue attempting to complete the task on his or her own. If the user was able to complete the task after an experimenter intervention, this was noted. Participants carried out the 17 tasks in identical, sequential order, providing verbal protocol feedback throughout the session. After completing the last task, all participants completed a user satisfaction questionnaire, provided debriefing comments and feedback, and then received a software gratuity. The sessions lasted approximately 1.5 to 2 hours and participants were run singly per session.

RESULTS

Task Success, Task Time and Time Estimates

The overall success rate for the seventeen tasks was 89.12% when the experimenter provided assistance to participants (much as would occur if customers called a family member, friend or Customer Service). If no experimenter assistance had been provided, this user group would have completed only 59.53% of the tasks. On average, the experimenter had to intervene and provided assistance 4.76 times per task, across all six participants.

The overall average task time was 203 seconds (st. dev .= 100 s). Task times ranged widely, from an average task completion time of 386 seconds for the task, "Add a 2nd account to the desktop machine" to an average of 86 seconds to "Read and close an email". Tasks taking longer than 4 minutes, on average, included Adding an account, Personalizing the home page, Adding a buddy and sending an IM, Playing media (music and video), and Adding Holidays to the calendar.

User Satisfaction

Users rated their overall satisfaction with the software at the end of completing the 17 tasks. Using a scale of 1=disagree, 5=agree, users rated the browser on a variety of dimensions that have been well validated across thousands of users as indicative of useful and usable software. Average satisfaction ratings are provided in Table 1 below. As suspected, despite the fact that success rates without experimenter intervention were quite low, on average (less than 60% of the tasks could be accomplished without the experimenter's assistance), users rated the browser quite highly on most of the dimensions of satisfaction. In fact, 17 out of the 19 questionnaire items were rated as above average in user satisfaction! Clearly, performance and satisfaction were

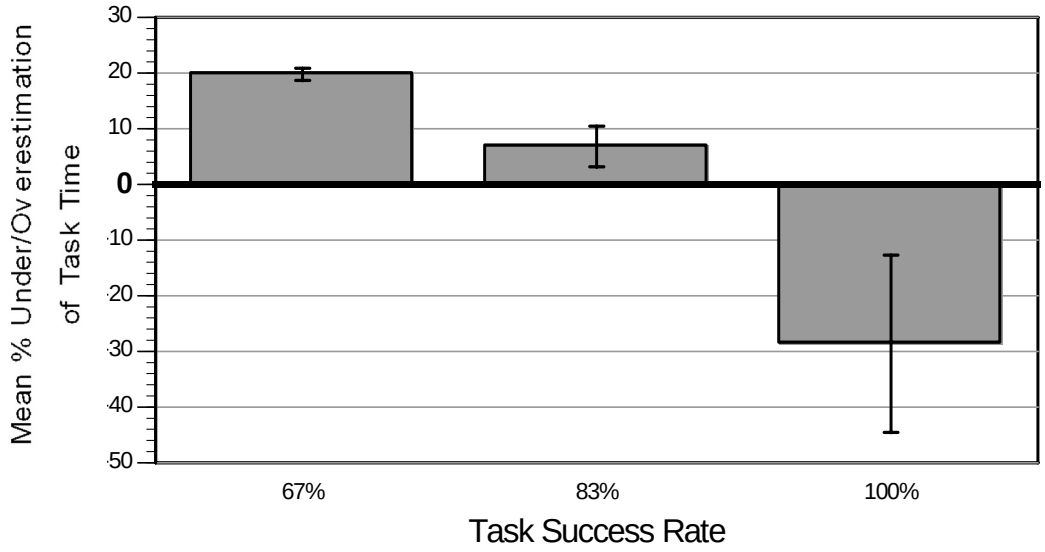

Figure 1. Over- and under-estimation of task times by task success rate (negative y-values are underestimates; positive y-values are overestimates).

not well correlated in this study, as is often the case in laboratory usability work.

Satisfaction Question	Average
I liked it.	3
I would use this software on a regular basis.	3.33
I would recommend this software to others.	3.5
The purpose of the software was clear.	3.5
Right when I started, I knew what I could do.	2.33
It was easy to get where I wanted to go.	2.67
Each area was clearly marked.	2.83
This software uses cutting edge technology.	3.67
This software provides valuable information.	4
This software provides detailed interaction.	3.83
This software has appealing graphics.	3.5
This software uses appealing audio.	3.5
This software is timely (or up-to-date).	4.167
This software is easy to use.	2.5
This software provides a shared experience.	3.67
This software is personalized/customizable.	4
This software feels unique (or different).	3.83
This software feels familiar.	3.17
This software is responsive (not too slow).	4
Overall Average:	**3.42**

Table 1. User Satisfaction Ratings and Comments

Time Estimation

In addition to collecting overall task times, we asked participants to estimate how long each task took upon completing that task (participants were not immediately told when they failed a task, in order to allow task time estimates if they thought they were finished). With only six participants, an interesting significant effect emerged, paralleling the Ziegarnik effect. The results showed that, indeed, participants reliably underestimated tasks with high success rates (% of participants completing the task), and reliably overestimated the lengths of tasks that had lower success rates. A multiple linear regression of participants' average over or underestimations against the tasks' success rates (% of participants completing a task) showed a significant linear relationship, $F(1,16)=6.08$, $p=.03$. In other words, as success rates for tasks went down, the estimation of the time it took to complete tasks increased reliably. Hence, task time estimation was an interesting "implicit" measure for usability, apparently tied to user satisfaction/frustration with the time it took to carry out browser tasks. Tasks for which participants overestimated the length should be considered high priority tasks to make more usable by the browser design team. A summary of the relationship between the actual versus estimated task time findings is shown in Figure 1.

DISCUSSION

This study showed that despite poor task success rates when using a prototype browser for typical web tasks, users tend to rate the browser's ease of use well above average. This often puts the usability engineer in a difficult position in terms of relaying the study results back to the team. There is good news, however , as the results of this user study also revealed that subjective duration estimates could provide an implicit measure related to the success of a user interface design for a given task. As a task becomes more difficult (perhaps due to an interface design), participants will likely overestimate how long that task takes. In contrast, if participants can complete the task, either with minor assistance or on their own, they are more likely to underestimate how long that task took in comparison to the actual task time. This intriguing result reveals

something akin to the modified Ziegarnik effect described by Weybrew (1984). In addition, participants do not necessarily know ahead of time which direction the experimenter expects the time estimates to go, and hence may not "bias" their reported estimates toward the positive end of the scale, as so often happens during lab studies using satisfaction measures in questionnaires [3]. In fact, after detailed experimenter questioning during study debriefing interviews, only one participant of the six thought that the time estimates might have had something to do with task success.

A few things remain unclear with regard to why a reliable Ziegarnik effect was found in this study. Did the effect have to do with the fact that participants in this study could not complete the less successful tasks, or did it have to do with the large number of interventions that likely accompanied the more difficult and less successful tasks? Van Bergen (1968) reviewed early studies of the Ziegarnik effect that partially address these concerns. For example, she reported that participants that are highly motivated to complete tasks correctly would most likely get the Ziegarnik effect. A myriad of other factors could contribute to this effect, such as anxiety, increased demands on limited attentional resources, etc. Variations on the early research by Ziegarnik showed that the effect was primarily due to the lack of completion of a motivating task, not just the interruption itself. Van Bergen also compared studies with many versus few interruptions/incompletions because she worried that if participants received too many interruptions, they would begin to expect them and place less importance on the primary task. However, she found little difference between the two groups of findings (most obtained the Ziegarnik effect). On the other hand, Fortin & Masse (2000) demonstrated that the expectation of interruptions and the "wait period" most heavily influenced overestimation. Further experiments are currently underway in an attempt to isolate these phenomena in a more rigorous experimental lab situation, using the subjective duration assessment metric.

BIBLIOGRAPHY

1. Fortin, C. and Masse, N. (2000). Expecting a break in time estimation: Attentional time-sharing without concurrent processing. *Journal of Experimental Psychology: Human Perception and Performance, 26*(6), p. 1788-1796.
2. Fortin, C., Rousseau, R., Bourque, P. & Kirouac, E. (1993). Time estimation and concurrent nontemporal processing: Specific interference from short-term-memory demands. *Perception and Psychophysics, 53*, 536-548.
3. Nielsen, J., and Levy, J. (1994). Measuring usability - preference vs. performance. *Communications of the ACM* **37**, 4 (April), 66-75.
4. Penney, T.B., Gibbon, J. & Meck, W.H. (2000). Differential effects of auditory and visual signals on clock speed and temporal memory. Journal of Experimental Psychology: Human Perception and Performance, 26(6), 1770-1787.
5. Van Bergen, A. (1968). Task Interruption. North-Holland Publishing Co., Amsterdam.
6. Weybrew, B.B. (1984). The Zeigarnik phenomenon revisited: Implications for enhancement of morale. Perceptual and Motor Skills, 58, p. 223-226.
7. Ziegarnik, B. (1927). Uber das Behalten von erledigten und unerledigten handlungen. Psychologische Forschung, 9, 1-85.